［サービスエクセレンス］

⑥サービスの公平性に配慮したデータ分析技術

神嶌敏弘 | 産業技術総合研究所

公平性に配慮したデータ分析

シェアリングなどの各種のサービスを運用するとき，家屋や自動車を貸与するかどうかなど各種の決定が行われ，そして，この決定にはデータ分析技術が活用されている．この決定を実社会で行うにあたっては，当然ながら各種の社会規範に配慮しなくてはならない．そうした規範の１つとして，性別や人種など生得的地位を各種の決定過程に利用しないという意味での公平性がある．そこで，本稿ではこの公平性に配慮したデータ分析技術である公平性配慮型データマイニングについて紹介する．データ分析において生じた公平性の問題事例を紹介した後に，これらの問題の技術的解決を目指して研究されている公平性配慮型データマイニングを紹介する．

データ分析での公平性の問題事例

データ分析技術は与信，採用，保険，裁判などの重要な決定にも利用されるようになってきている．このとき，社会的･法的な公平性への配慮，すなわち，性別や人種など生得的地位に対する差別が生じないようにすることは重要である．この点に関して懸念が示された，ネット広告配信と再犯リスクスコアの２つの事例を紹介する．

ネット広告配信の事例

まず，ネット広告配信における Sweeney による指摘を紹介する[1)]．読者は，多くの文書や Web ページから，必要な情報を見つけ出す情報検索サイトを毎日利用していることと思う．これらのサイトでは，キーワードに関連した項目に加え，そのキーワードに関連した広告も併せて表示される．Sweeney はこの広告が，人種に対する偏見に基づいている可能性について調査した．具体的には，マサチューセッツ州の新生児の記録から，アフリカ系とヨーロッパ系の間で偏りが大きい 2,000 種以上の名前を選び，これらの名前で検索サイトとニュースサイトで検索し，表示される広告を調査した．

Sweeney は，米国の各州で公開されている逮捕歴の情報などを検索する Instant Checkmate などのサイトに関する広告に注目した．図 -1（a）は，アフリカ系に多い名前“Latanya Farrell”で検索した場合に表示された広告の例である．１行目の広告は『Latanya Farrell は逮捕されたか？』という逮捕歴を示唆するような広告文になっている．一方で，ヨーロッパ系の名前“Jill Schneider”で検索した図 -1（b）では，２行目の『Jill Schneider を見つけました』のように，特に逮捕歴を示唆しない中立的な広告文であった．より詳しく調べると，実際のリンク先のサイトで逮捕歴があるか，また，その名前のレコードが存在するかに基づいて広告文が選択されているわけではなかった．アフリカ・ヨーロッパ系の区別と，広告文が中立かどうかの独立性を統計的に検定したところ，有意にアフリカ系で逮捕歴を示唆する広告文が多かったと報告している．

さらに，Instant Checkmate 社に対してインタビューによる調査を行ったが，単純に広告の収益率

を最大化するようなものを選択しており，恣意的な差別は認められなかった．このテンプレートは姓のみに基づいて選択しており，他の規準はないとのことであった．これは，データ分析技術自体には偏見はないが，データに含まれる社会の悪意が，意図せず反映されてしまった事例といえる．

再犯リスクスコアの事例

次に，データジャーナリズム NPO ProPublica による再犯リスクスコアに対する指摘を紹介する[2]．データジャーナリズムとは，データ分析を用いたエビデンスに基づくジャーナリズムで，記事とともに分析過程やデータをも公開している．

ここでいう再犯リスクスコアは，被告人が 2 年以内に再び犯罪を犯すかどうかの可能性を評価するものである．過去の裁判システムには人種に対する偏見があったとの反省にたち，エビデンスに基づく決定を重視するという方針で導入が進んでいる．この記事中でも指摘していることではあるが，スコアの導入自体ではなく，そのスコアに偏りがあることを問題視している点が重要である．このようなエビデンスを重視する方針がなければ，統計分析に基づいたこうした厳密な議論すらできなかったであろう．

本題に戻り，再犯リスクの予測の傾向が人種間で異なっているとの ProPublica の分析結果を紹介する．具体的には，実際には 2 年間に再犯しなかった人が，再犯すると誤って予測されてしまった割合は，アフリカ系が 45% であるのに対し，ヨーロッパ系では 23% であった．すなわち，アフリカ系の人について，実際には更生する人を再犯すると予測してしまいやすい．逆に，その後 2 年の間に実際に再犯しまった人を，再犯しないと誤って予測してしまった割合は，アフリカ系が 28% でヨーロッパ系が 48% であった．すなわち，実際には再犯するヨーロッパ系の犯罪者を見逃しやすい．ここで注意すべきは，全般的にはアフリカ系の場合の方が実際の犯罪率が高いので，アフリカ系の人を全般的に高リスクであると判定してしまうことを問題としているわけではない．大まかにいえば，予測には不確実性が必ず伴うが，この不確実性に人種間で差があるという指摘である．

予測アルゴリズムはデータ量が十分であるなど条件が整えば，こうした予測の不確実性には偏りが生じないように設計されている．しかし，実際にはこうした条件が十分に満たされない場合もあり，特に悪意ある操作をしなくても，この事例のような差が生じ得る．

(a) アフリカ系の“Latanya Farrell”で検索した場合

(b) ヨーロッパ系の“Jill Schneider”で検索した場合

■図 -1　人名で検索した場合に表示される Web 広告の例☆1

☆1 © 2013 Association for Computing Machinery, Inc. This translation is a derivative of ACM-copyrighted material. ACM did not prepare this translation and does not guarantee that it is an accurate copy of the originally published work. The original intellectual property contained in this work remains the property of ACM[1].

公平性配慮型データマイニング

前章のような問題に対処するため，データ分析の過程で，公平性に配慮するのが公平性配慮型データマイニング（fairness-aware data mining）である．ここでは，公平性についての形式的な規準を紹介したのちに，これらの規準を用いた分析タスクについて簡単に述べる．

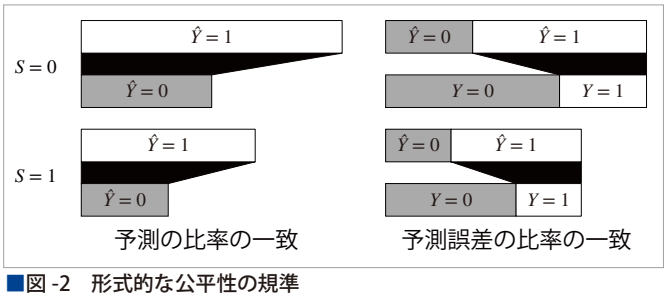
形式的な公平性の規準

計算機上で公平性を扱うための形式的な公平性の定義について説明する．準備として，目的変数 Y とセンシティブ特徴 S について述べる．目的変数は，分析の目的を表すもので，与信の可否や，就職における採用などを表す．公平性は，センシティブ特徴の表す情報について後述の規準を満たすように保証する．たとえば，与信や採用などの決定で，社会的公平性の観点からその関与を排除すべき対象者の性別や人種といった個人属性情報を，このセンシティブ特徴で表す．ここでは，Y も S もその値が，0 か 1 のいずれかである簡潔な場合について扱う．たとえば，採用の場合は $Y=1$ で，不採用なら $Y=0$ にする．また，センシティブ特徴についても $S=1$ なら男性，$S=0$ なら女性のようにする．Y については，さらに結果の予測結果を $\hat{Y}$，実際の値を Y として区別する．採用の場合でいえば，分析中のデータで採用の可否を表すのが Y，採用すべきかどうかの予測結果を表すのが $\hat{Y}$ となる．

この公平性は，センシティブ特徴を単純に分析で使わないだけでは保証できない．なぜなら，ほかの特徴との間に相関があれば，その特徴を通じてセンシティブな情報が結果に影響してしまうからである．たとえば，人種ごとにまとまった地域に住んでいることはよくあるため，たとえ人種という情報を直接的に使わなくても，住所の情報を用いて分析すると間接的に人種の情報を使ってしまうことになる．人種を直接的な理由とせず，地図上で赤枠で囲った特定地域の住人に銀行が貸し出しをしなかった過去の事例から，この現象を red-lining 効果[3]と呼ぶ．この red-lining 効果がデータ分析での公平性の技術的取り扱いを難しくしている．

形式的な公平性規準を 2 つ紹介する．1 つ目の公平性規準を説明するために，最初の広告配信の例を考えよう．この場合は，社会の偏見に基づく判定がデータに含まれていることが原因となっている．すなわち，アフリカ系の人に対して犯罪歴を示唆する広告文があるとクリックするという偏見のある判断がデータに含まれている．よって，データの判断を部分的に無視して，センシティブ情報が変わっても同じ割合で広告文を選ぶようにする形式的公平性を考える[3]．すなわち，センシティブ特徴 S がアフリカ系とヨーロッパ系とのいずれの場合でも，選択する広告文の比率を一定に保つようにする．これは，図 -2 左の $S=0$ と 1 の 2 つの場合の比率（図左の黒塗り部分）が一致するということである．参考までに，この規準を数式で表すと，これは $\hat{Y}$ と S の統計的独立性 $\Pr[\hat{Y}, S] = \Pr[\hat{Y}]\Pr[S]$ にあたる．

もう 1 つの公平性規準を，再犯リスクスコアの例に基づいて紹介する．この場合は，再犯をするかどうかは客観的な基準に基づいているのでデータには偏りは生じない．ここでの問題は，データ量が十分ではないなどの理由により予測結果に偏りが生じているということである．再犯リスクでは，スコアによるリスクの高低を $\hat{Y}$，実際に 2 年に以内に再犯したかを Y で表すことになる．このとき，実際の結果 Y に対して，予測結果 $\hat{Y}$ がどれくらい外れてしまうのかを，センシティブ特徴の値によらないように調整する．これは，図 -2 右のように，データ Y



■図 -2　形式的な公平性の規準

と予測 $\hat{Y}$ の比率（図右の黒塗り部分）が，センシティブ特徴 S の値が 0 であっても 1 であっても同じになるようにする[4]．この規準は数式で表すと，Y が与えられたときの $\hat{Y}$ と S の独立性 $\Pr[\hat{Y}, S|Y] = \Pr[\hat{Y}|Y]\Pr[S|Y]$ となる．なお，以上 2 つの規準は同時には達成できない場合があることが知られている．

分析タスクの分類

公平性配慮型データマイニングの分析タスクは，大きく不公平発見（unfairness discovery）と不公平防止（unfairness prevention）に分類できる．不公平発見では，分析結果に不公平なものが含まれているか，また含まれているとすればその結果を抽出する．不公平防止とは，不公平な分析結果が生じないようにしつつ，クラス分類や回帰といった分析を行う手法である．最後に，これらについて簡潔にまとめる．

不公平性検出の例として，データ分析における公平性を扱った最初の研究を紹介する[5]．これは相関ルールというものを対象としている．相関ルールは，データがある条件を満たす場合（たとえば，前科がなく 40 歳以上）は，目的変数がほぼある状態になる（たとえば，再犯はしない）というデータ中の関係性を表すものである．この相関ルールの中に，センシティブ特徴の値が目的変数に大きく影響するものがあるかを検査し，それらを列挙する問題を提唱した．

不公平防止タスクについては，これらの形式的公平性の基準を満たすような制約の下で，予測精度を最大化するようなアルゴリズムが研究されている．このタスクの最初の研究は単純ベイズと呼ばれるクラス分類の手法を対象としていた[3]．通常のクラス予測の規則を獲得したのち，センシティブ特徴によって生じる分類の偏りを補正した．クラス分類以外にも，回帰などの多様な予測タスクがあるが，これらについても研究が現在では広がっている．筆者も，利用者の嗜好に合うであろうものを予測する推薦タスクについて研究している[6]．

以上，データ分析分野における公平性について解説した．今年 2018 には，新たな国際会議 Conference on Fairness, Accountability, and Transparency[☆2] も設立された．こうした学会などを通じ，データ分析における公平性の議論が今後も深まっていくであろう．

参考文献

1) Sweeney, L. : Discrimination in Online Ad Delivery, Communications of the ACM, Vol.56, No.5, pp.44-54 (2013).
2) Angwin, J., Larson, J., Mattu, S. and Kirchner, L. : Machine Bias, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3) Calders, T. and Verwer, S. : Three Naive Bayes Approaches for Discrimination-free Classication, Data Mining and Knowledge Discovery, Vol.21, pp.277-292 (2010).
4) Hardt, M., Price, E. and Srebro, N. : Equality of Opportunity in Supervised Learning, In Advances in Neural Information Processing Systems 29 (2016).
5) Pedreschi, D., Ruggieri, S. and Turini, F. : Discrimination-aware Data Mining, In Proc. of the 14th ACM SIGKDD Int'l Conf, on Knowledge Discovery and Data Mining, pp.560-568 (2008).
6) Kamishima, T., Akaho, S. and Sakuma, J. : Recommendation Independence. In Conf. on Fairness, Accountability and Transparency, Vol.81 of PMLR, pp.187-201 (2018).

（2018 年 1 月 12 日受付）

☆2 https://fatconference.org/

神嶌敏弘 mail@kamishima.net
1992 年京都大学工学部情報工学科卒業．1994 年同大学院工学研究科情報工学専攻修士課程修了．2001 年博士（情報学）．1994 年電子技術総合研究所入所．2001 年産業技術総合研究所へ再編．推薦システム，データマイニング，機械学習に関する研究に従事．